The BADC Text File
Guide for users, and producers.

Sam Pepler, BADC, 2008-09-23

Verson	Who	Date	Notes
0.1	Sam Pepler	2008-09-23	Initial version make from dev document.
0.2	Graham Parton	2008-10-14	Addition of flag_values and flag_meanings
0.3	Graham Parton	2009-04-23	Cleaned up document and updated conventions used
0.4	Graham Parton	2010-04-22	Minor corrections

About this document

This document is aimed at people developing or producing data files in the BADC text file format. It describes the structure of the format and the rules for its content.

Format History: An Alternative to NASA-Ames

The BADC has used NASA-Ames formatted data for many years. NASA-Ames was devised primarily as a format for aircraft observations, but can be adapted for many atmospheric observation data. However, NASA-Ames is complex and confusing for users. Users tend to strip the header off and import the text file into Excel. The metadata is generally not used in its machine readable form, but is simply read by the researcher. Also much effort is expended supporting data producers in the creation of NASA-Ames files. The format is seen by producers as complicated and it can't be done simply from spreadsheet packages like Excel. Additionally, the metadata fields offered by NASA-Ames are fixed and inflexible.

Model data stored at the BADC often uses the NetCDF format with CF conventions. This provides a format framework with good flexible metadata. The format can be read by a number of analysis programs including FORTRAN, Matlab and IDL. It is however difficult for a researcher with little technical knowledge to use.

To solve these problems a new file format was developed to bring the advantages from the NetCDF file format into a simple text file. The approach was to use metadata conventions on top of comma separated values files (CSV) as produced by applications like Excel.

Metadata conventions for CSV

CSV nomenclature

- A line is a single CSV record ending in a line feed (i.e. \r\n)
- An entry is a single comma delimited field

A BADC text file contain 3 sections

- File type identifier
- Metadata
- Data

File type identifier

The first metadata line in the file should be the Conventions line. This aids recognising the file type. This is given as shown below to conform to the CF conventions and is the only metadata field that is capitalised. All others that follow this line are in lower case.

Conventions,G,BADC-CSV,1

Metadata section

The all metadata entries are of the format:

<label>, <column ref>, [<value>, <value>, …]

<label> is a metadata tag which may be an item form the list of controlled metadata items in the appendix to give greater conformity to various metadata standards, or may be one generated by the user. **Note that items in the controlled list have special meanings and should not be used other than as stated.** All words within a metadata tag should be joined with the underscore ("_") and not whitespace and should be entirely lower case.
<ref> is the column reference to which the metadata applies. A "G" indicates that the metadata applies globally. This allows reference to variables and the data file in the same intuitive manner as NetCDF. See below regarding column referencing.
<value>, … is the set of one or more comma separated values associated with the metadata line. The number of values depends on the metadata item, for example "data_valid_range" needs a start and end date/date-time, while "coordinate_variable" can be used to indicate a column as being a coordinate variable while its associated values are optional.

To aid readability it is permissible to have repeat metadata tags to allow values to be split over more than one line. See example 2 below where this has been done for some comment lines.

Note on required fields: the controlled metadata items listed in the appendix include items that are mandatory for all BADC-CSV files and are indicated by the comment “Basic” in the column entitled “Needed at Compliance level”. Those fields listed as “Complete” should be provided where possible to provide compliance with various metadata standards. Comments in braces indicate additional requirements for the metadata field for the file to follow certain standards.

Note on column referencing: to link metadata elements to the relevant data elements references are used with each metadata element and are listed along the top of each column as the first line following the “data” line in the data section (see next section). These links can either be numerical as in the first example file or text based as in the second example file. “G” should not be used as a column reference as this is reserved to indicate where metadata fields apply globally. Finally, each column should have a unique column reference and metadata entries must either apply globally or reference only one column. If metadata information is identical for two or more columns the entries should be repeated in the metadata with appropriate column referencing.

Some dataset producers may wish to use column numbering as a useful reference and then provide additional metadata tags indicating the mapping of these numbers to a text description, e.g.

column_headings,G,<1st col. name>,<2nd col. name>,<3rd col. name> etc

This could be placed immediately above the “data” line which marks the top of the data section to aid locating in a file with extensive header information.

Unknown entries within the metadata values should be given by the word “unknown” and not be omitted, left blank or indicated by “?” or other word/symbol. This helps to provide confidence to file users that the value really was unknown rather than having been accidentally omitted or missed by some generating script.

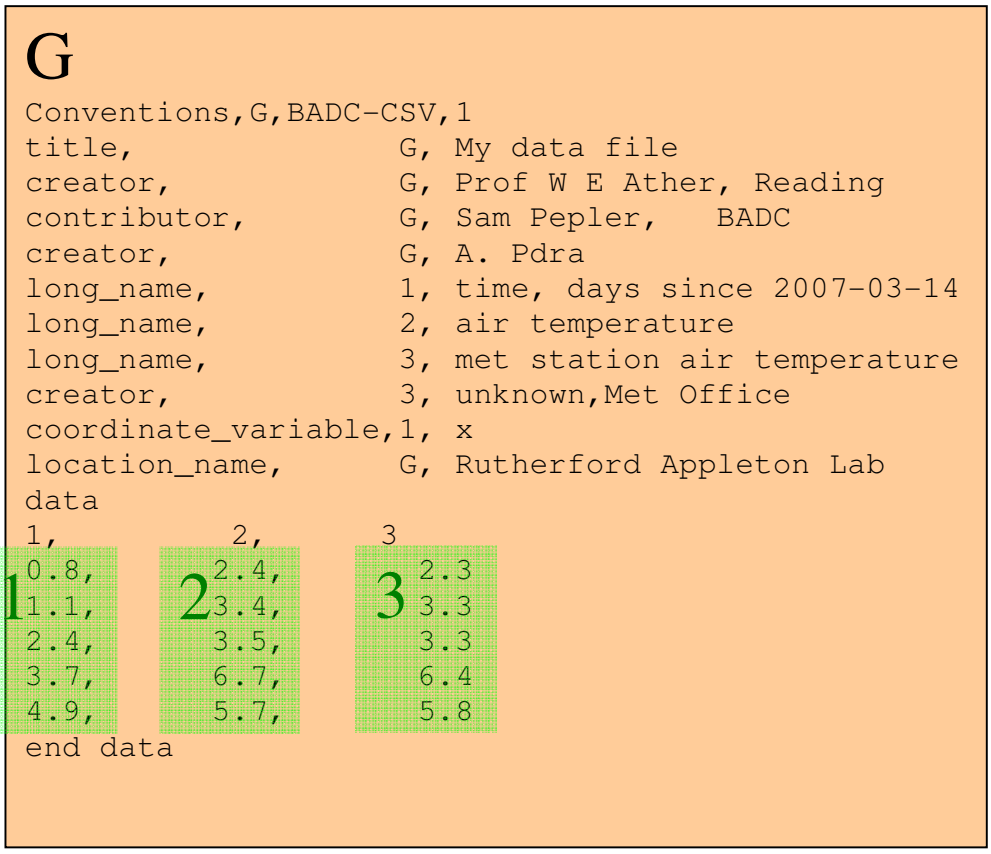
Data section

The data section consists of a record with a single “data” entry, followed by a line of the column references and then the data records. The end of the data records is indicated by an “end data” entry. The end data entry is included to flag partial files. Both “data” and “end data” are to be given in lower case.

data
<column references>
<data lines>
end data

Examples

The examples below illustrated the concepts given above. The orange box labelled G indicates the extent of the global attributes (i.e. to the entire file), while the referenced green boxes link to the relevant metadata entities. Entries have been spaced out below to aid readability, but generally use of additional white space is not encouraged as this is superfluous to requirements within csv files.



This example show a file, titled “My data file” created by Prof W E Ather at Reading. The data have been prepared by A. Pdra, but the some metadata was added at the BADC by Sam Pepler. There are three variables in the file: time (column 1), air temperature (column 2) and met station air temperature (column 3). The met station air temperature was created by an unknown person at the Met Office. The time variable is flagged as the coordinate variable (and marked as suitable for plotting on the x axis of a graph). The units of the time variable are given in days since 2007-03-14 . The data in the file was measured at the Rutherford Appleton Lab.

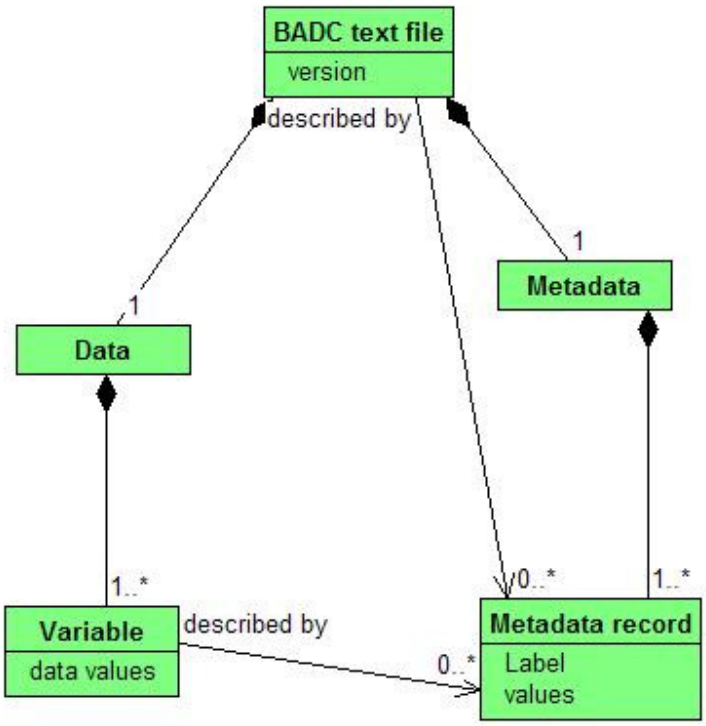
```
G

Conventions,G,BADC-CSV,1
title,                 G, My data file
creator,               G, G Parton, CEDA
contributor,           G, Sam Pepler,   BADC
creator,               met_temp, S Aylingby, CEDA
long_name,             time, time, days since 2007-03-14
long_name,             temp, air temperature
long_name,             met_temp, met station air temperature
creator,               met_temp, unknown,Met Office
comments,              met_temp, measured using a thermometer
comments,              met_temp, the instrument_materials
comments,              met_temp, field details the main
comments,              met_temp, material of the instrument
comments,              met_temp, only
instrument_materials, met_temp, glass and mercury
coordinate_variable,1, X
location_name,       G, Rutherford Appleton Lab
data
time,      temp,       met_temp
0.8,       2.4,        2.3
1.2,       3.1,        3.2
2.4,       3.5,        3.3
3.7,       6.7,        6.4
4.9,       5.7,        5.8
end data
```


In this example the column references have been replaced by text references and a user defined metadata tag “instrument_materials” has been added to allow additional details to be provided with the file. Comment lines have been added to let others know what this additional metadata tag represents and to aid readability and to ensure that all text is on one screen the comment has been split into 4 consecutive lines.

Data model

The diagram below shows the data model for the file structure. It says a BADC text file consists of data and metadata. The data consist on variables and the metadata consists of metadata records. All these classes are described by the metadata records.



Levels of compliance

There are five levels of compliance that can be usefully identified to provide a checking framework.

- **CSV**: The file should conform to Excel dialect CSV file format. The rules for this are fairly clear and most applications and programming languages already support it.
- **Structure**: Data and Metadata sections exist
- **Valid metadata**: Metadata has right number of values and refers to legal objects.
- **Basic**: Parameter names for all columns exist. This provides a file with the same information numbers and column headings. The basic structure of the file is correct. This level requires valid metadata.
- **Complete**: Mandatory metadata exists. Metadata should exist for some items. Requires basic compliance.
- **Standardised**: Metadata values for appropriate is from standard list. Requires complete compliance.

Appendix: Metadata values

Number of values lists the minimum to maximum number of variables that can be supplied. Information in braces following the number details what each of the given values should be in order or appearance. Optional values are given in square braces.

The order of metadata given below group basic mandatory fields first, followed by those required for a file to be considered complete with regards to metadata requirements and then optional controlled metadata tags. The order within each group is (with the exception that Conventions must appear as the first line) a guide to the relative location of the various metadata tags within the metadata section, e.g. it is suggested that a file would appear more logically ordered if long_name preceeds any other metadata entries for a given column (e.g. standard_name).

Metadata label	**Description**	**Can be applied to**	**Number of values**	**Needed at Compliance level**	**Metadata convention**
Conventions	Conventions used in metadata	G	2 (BADC-CSV, version number)	Basic	CF
long_name	Text description of the variable and unit	C	2 (name, unit)	Basic (Standardised units)	NA, CF
coordinate_variable	Flag for coordinate variables, optionally a plotting axis suggestion and the name of the coord ref system	C	0-2 (xyt, ref system)	Basic (from list for standard)	CSML, CF, ISO191115
creator	As DC Creator	G, C	1,2 (name, inst)	Complete	DC
source	Name of tool used in the production of the data	G, C	1	Complete	NA, MOLES

observation_station	Name of the observation station or platform used	G, C	1	Complete	ISO19115, MOLES
activity	Name of the activity sponsoring the collection of the data	G, C	1	Complete	NA, MOLES
feature_type	Profile, point series, trajectory, point collection	G	1	Complete, (from list for Standardised)	CSML
location	Location valid for data.	G, C	1 – name of location or WKT formatted 2 – (lat, long) 4 – bounding box, (?,?,?,?)	Complete	ISO19115
date_valid	Date valid for data	G, C	1,2 (start, end)	Complete (ISO8601 date and (optionally) for Standard)	ISO19115
last_revised_date	Date the data, file or metadata was last changed.	G, C	1	Complete (ISO8601 date and (optionally) time for Standard)	DC, NA
history	Free text description of the file history	G, C	1	Complete	CF
standard_name	Name from CF standard name list	C	3 (name, unit, standard)	Complete, (CF name for Standardised)	CF

title	As DC title	G	1		DC, CF
comments	Any text associated with data	G, C	1		NA
contributor	As DC	G, C	1,2 (name, inst)		DC
height	Height valid for data	G, C	2-4(min height,[max height], unit,[datum])	(Standardised units)	
reference	Bibliographic reference	G, C	1		DC
rights	As DC rights. Conditions of use for the data	G, C	1		DC
valid_min	Values below min should be interpreted as missing.	G, C	1	(float for standard)	NA, CF
valid_max	Values above max should be interpreted as missing.	G, C	1	(float for standard)	NA, CF
valid_range	Values outside this range should be interpreted as missing.	G, C	2 (min, max)	(float for standard)	NA, CF
type	The value interpretation for the variable – char (default), int, float)	C	1	Complete (from list for standard)	NetCDF
cell_method	As CF.	C	1		CF
add_offset	Offset for data values. As CF	C	1	(float for standard)	CF

scale_factor	A scale factor for the data. As NA	C	1	(float for standard)	NA
flag_values	As CF: values used for flag table in data	C	List		CF
flag_meanings	As CF: meanings for each flag_value	C	List		CF

Notes:
Date and Times
All dates should be given in the YYYY-MM-DD format, while times are represented as hh[:mm:ss.sss] down to the required resolution. Fractions of a second can be given if appropriate to the degree required.
When the date and time elements are to be given together then they should appear separated by a space i.e. YYYY-MM-DD hh:[mm:ss].

E.g. if jus the hours of observation are to be given for data taken at the synoptic hours (00,06,12 and 18 UT) for the first week in the year then the data_valid field could be given as:

date_valid,G,2006-01-01 06, 2006-01-08 18